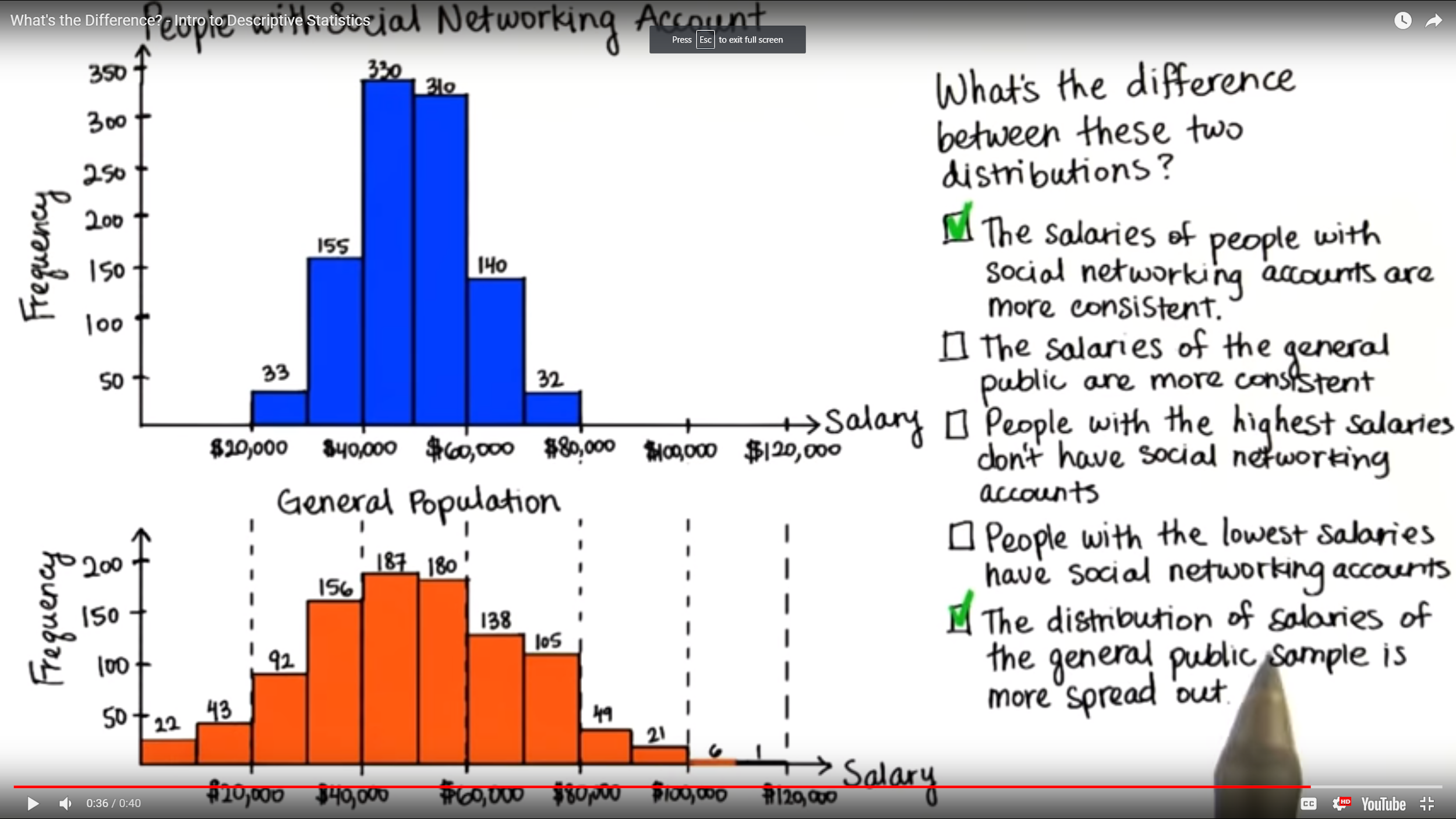
BS – Important Concepts Variability – Lessons 13, 14



1. Quantify spread: range

→ range = maximum value – minimum value

→ range sometimes changes when we add new data to the data set

→ we loose details on our range because it is only based on two scores, the most extreme in the distribution, are likely not very representative of the rest of distribution;

→ may not accurately represent the variability of data, especially in the presence of outliers;

→ don't take all data into account

1. Outliers:

→ increase significantly the range of our data;

→ increase the variability;

→ statisticians cut the lower 25% and the upper 25% tail of the distribution => they consider the values in the middle

→ Outlier < Q1 – 1.5(IQR)

→ Outlier > Q3 + 1.5(IQR)

1. Quartile:

→ The first quartile, Q1, is the point where 25% of the distribution is below that point, and 75% of the data is above that point.

1. Interquartile Range (IQR):

→ IQR = Q3 – Q1

→ about 50% of the data falls within the IQR

→ the IQR is not affected by every value in the dataset

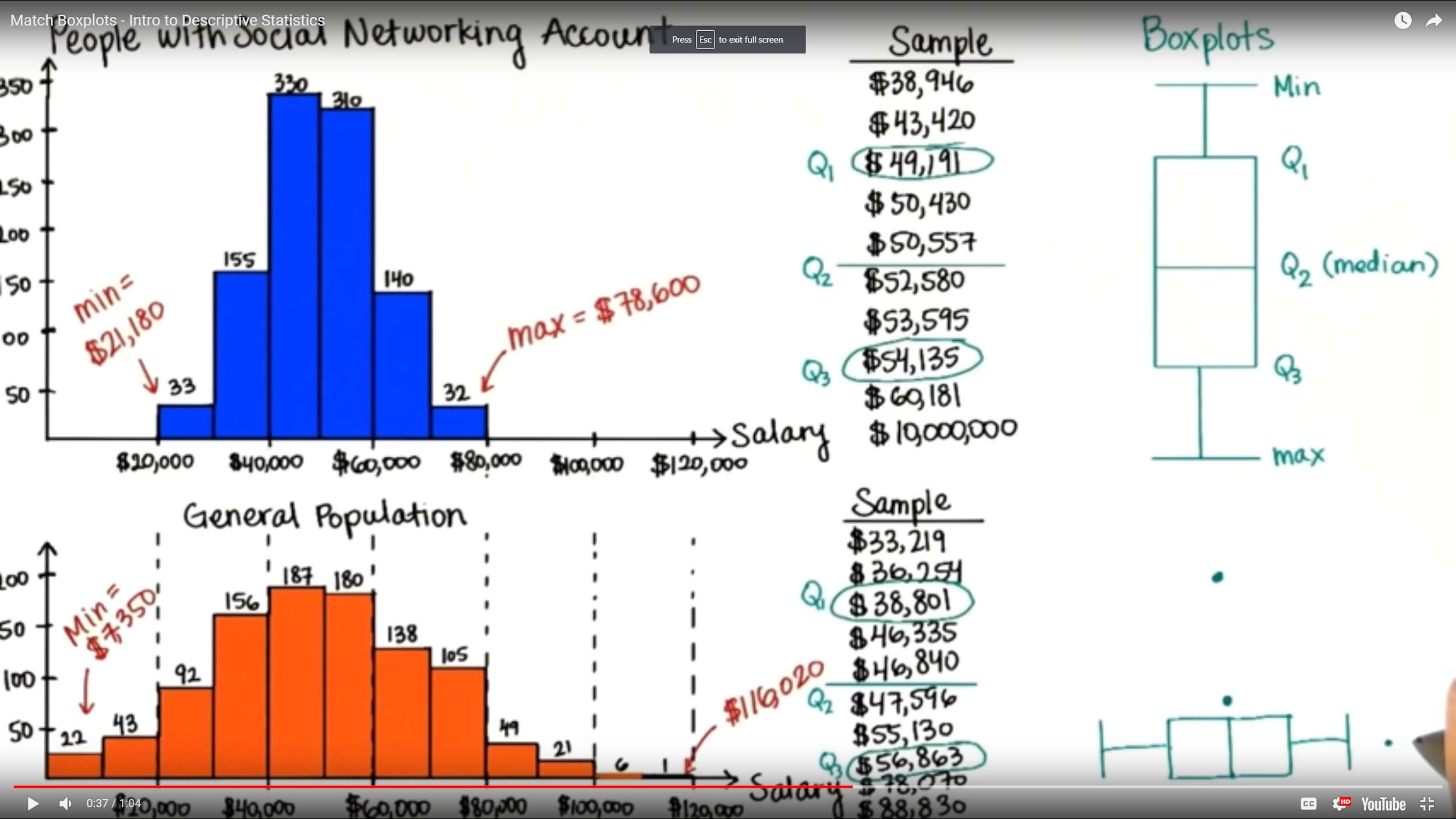
→ the IQR is not affected by outliers

→ the mean is not always between Q1 and Q3 when there are outliers, although usually it is in normal distributions

→ don't take all data into account

→ it doesn't tell us as much as we would like to know about the dataset, there can be different distributions with the same IQR

1. Match Box plots



1. Measure Variability:

→ We need one number that describes the spread of data that takes all the data into account

→ find the average distance between each data value and the mean (is in the center of the data set and takes all data into account);

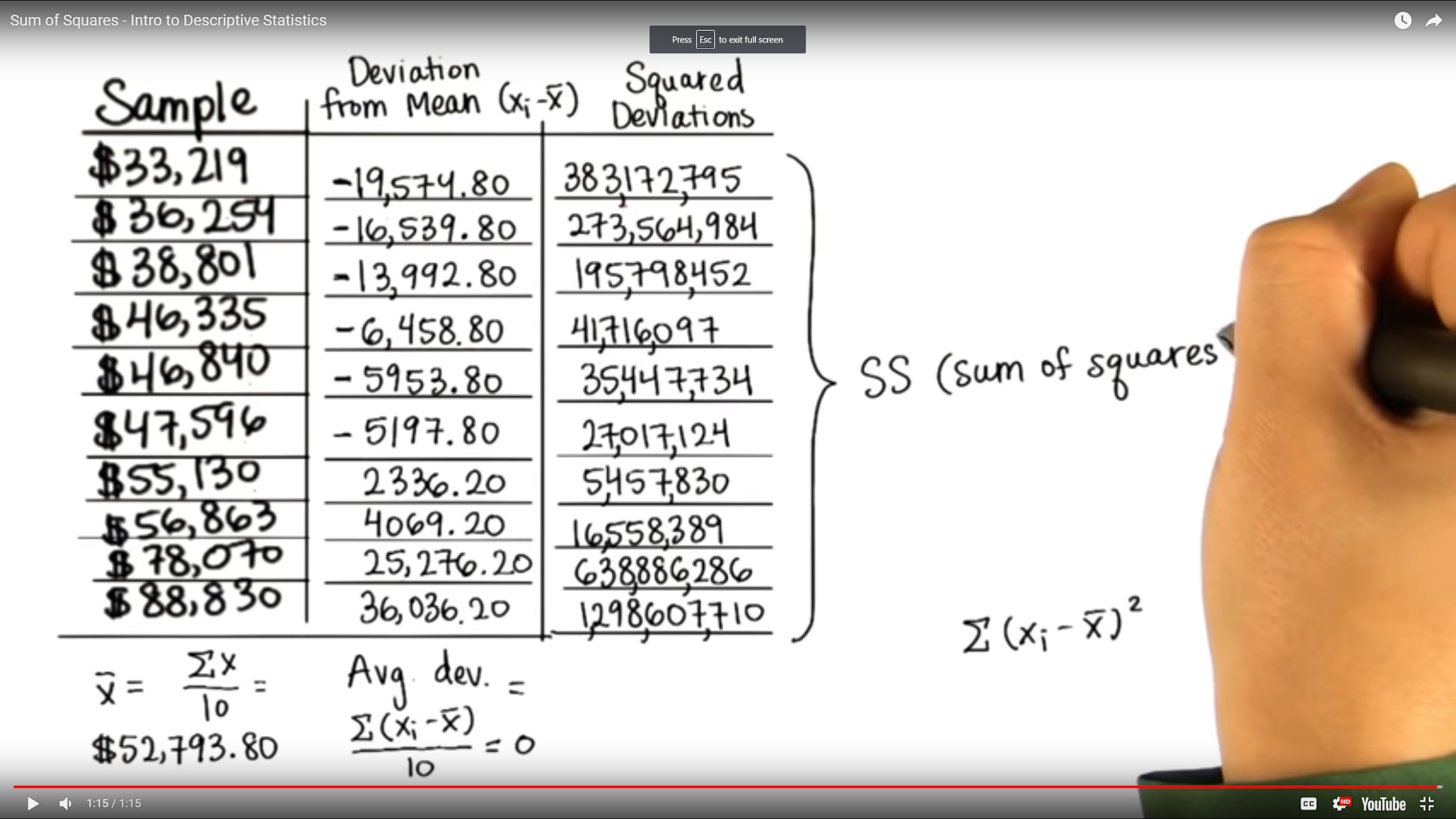
→ Deviation from the Mean: (Xi – X bar) and if we calculate the Average Deviation: sum(Xi – X bar) = 0

→ the sum of all deviation from them Mean = sum(Xi – X bar) = 0

→ the Average Deviation cannot be our measure of spread because is 0 => have to get rid of negatives, they are canceling out the positives

→ we care how far is the data from the mean → square each deviation

1. SS (sum of squares)



1. Variance (Average Squared Deviations) = (Sum of Squares)/n

→ mean of squared deviations

→ sum of squared deviations divided by n

1. Standard Deviation = sigma

→ distance between each value and mean and squared them = area of every square whose side length is the distance between each value and the mean = SS

→ distance = Xi – Xbar

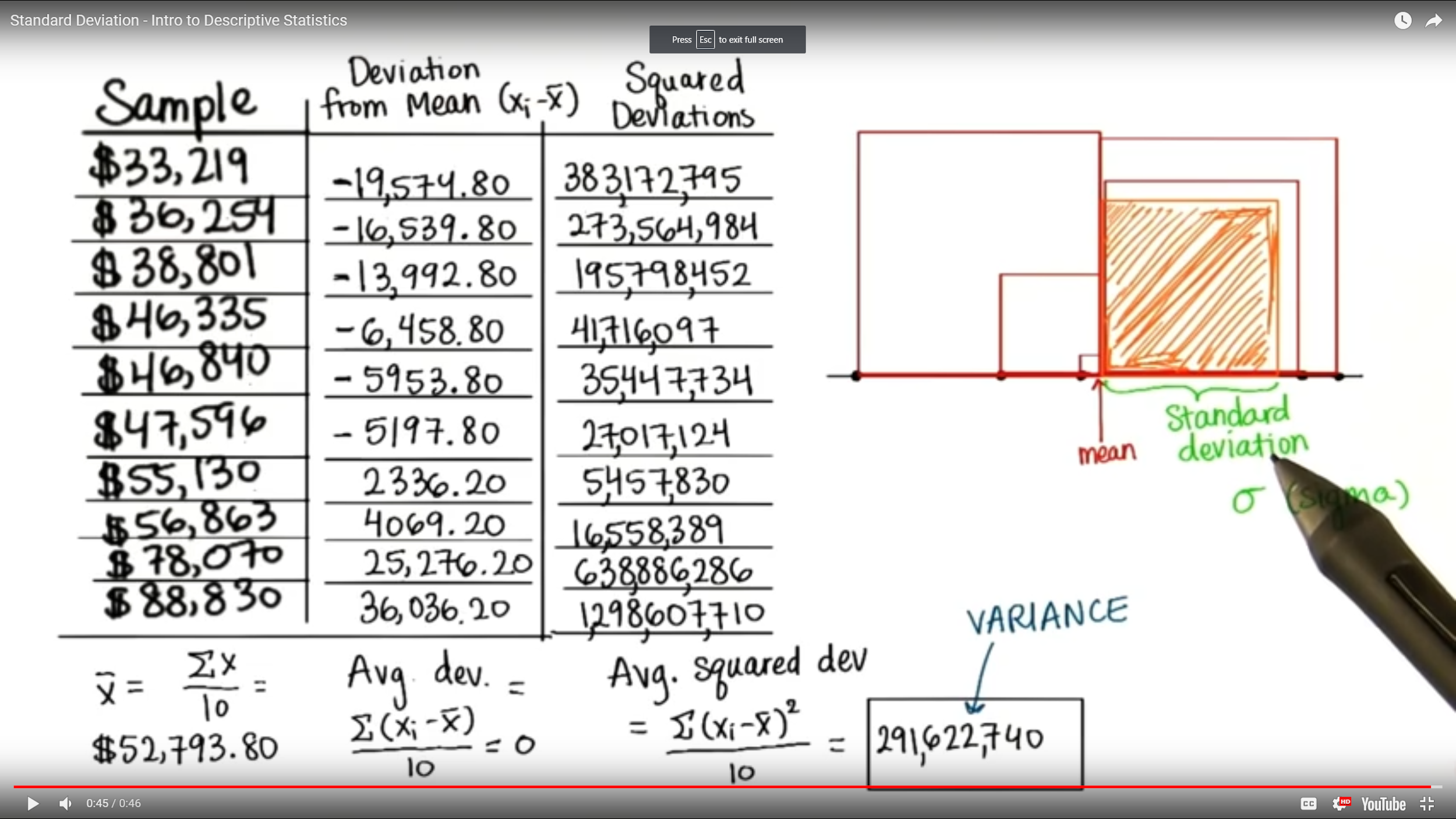
→ area = (Xi – Xbar)^2 => to convert it to one dimension we have to take the square root of it => we end up with just the length of one of the sides = standard way of measuring the variability = Standard Deviation

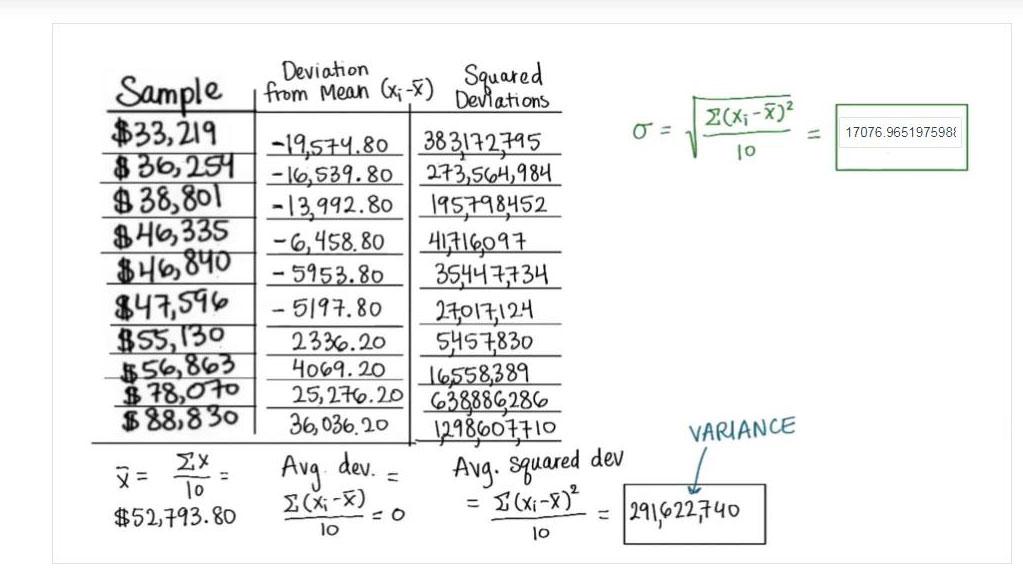
→ Standard deviation: find each deviation from the mean, squared it, find the average area of the squares, take the square root to find just the length of the side

→ the most common measure of spread

→ Square root of average squared deviations

→ Square root of ((sum of squared deviations) divided by n)





1. With a normal distribution the Standard Deviation has great properties:

→ 68% of all data falls within 1 standard deviations of the mean;

→ 95% of the data falls within 2 standard deviations of the mean

1. Bessel's Correction (s) = Sample Standard deviation

→ the variability in the sample is less than the variability of the entire population =>Bessel's Correction is used to correct for this

→ this will make the original standard deviation and variance bigger

→ we use it to approximate the true population standard deviation

→ used to estimate the Standard Deviation of the population

